university of groningen

University Medical Center Groningen

University of Groningen

The Statistical Physics of Learning (in a nutshell)

Biehl, Michael

DOI:
10.1007/s10618-017-0506-1

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2018

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Biehl, M. (2018). *The Statistical Physics of Learning (in a nutshell): News from the stoneage of neural networks*. 23-23. Abstract from Mittweida Workshop on Computational Intelligence , Mittweida, Saxony, Germany. https://doi.org/10.1007/s10618-017-0506-1

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

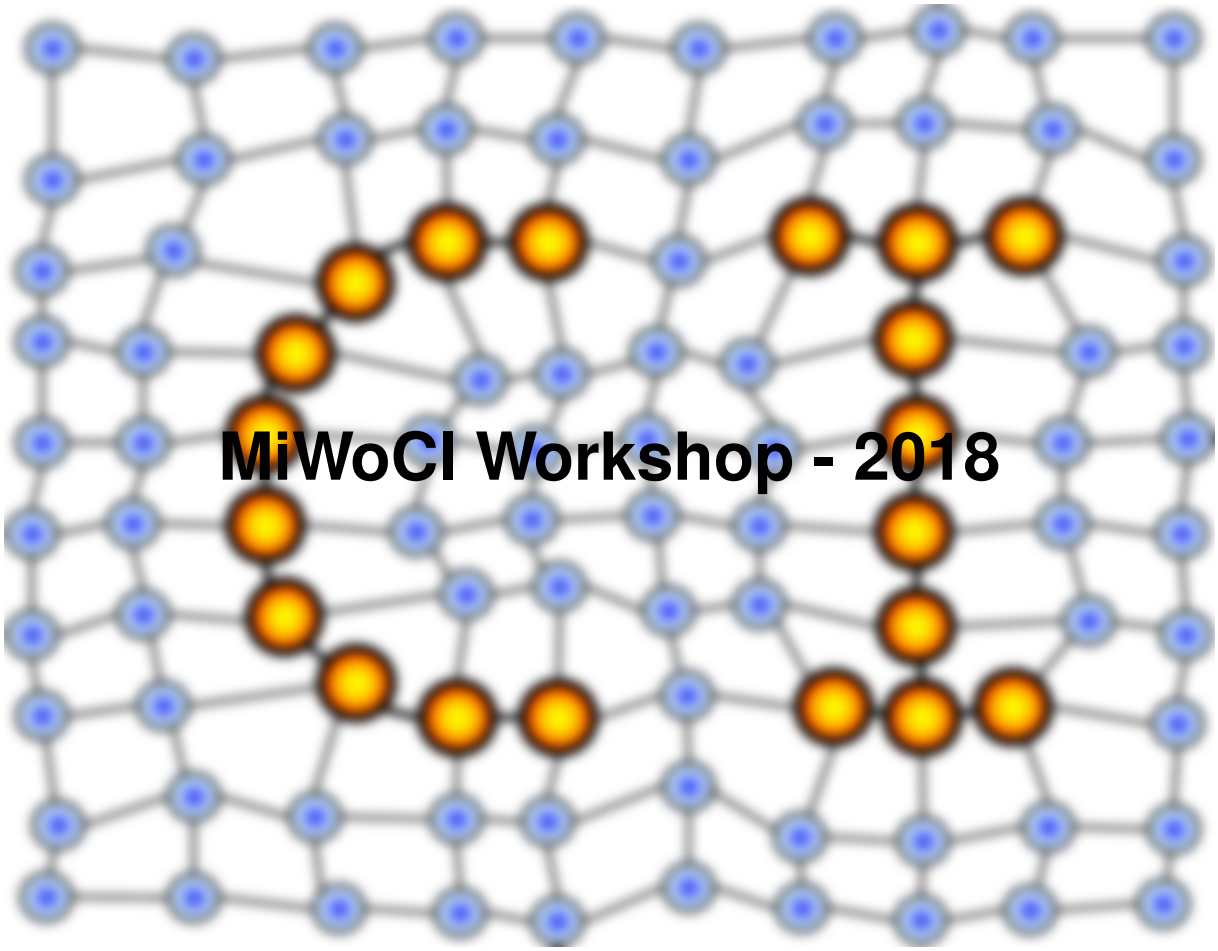
The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 05-05-2023

MACHINE LEARNING REPORTS



MiWoCI Workshop - 2018

Report 01/2018
Submitted: 23.06.2018
Published: 25.06.2018

Frank-Michael Schleif[1,2*,3*], Thomas Villmann[2] (Eds.)
(1) University of Applied Sciences Wuerzburg-Schweifurt, Sanderheinrichsleitenweg 20, 97074 Wuerzburg, Germany (2) University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany (3) University of Birmingham, School of Computer Science, Edgbaston, B15 2TT Birmingham, UK

Contents

Impressum

Publisher:	University of Applied Sciences Mittweida Technikumplatz 17, 09648 Mittweida, Germany
Editor:	Prof. Dr. Thomas Villmann Prof. Dr. Frank-Michael Schleif
Techical-Editor:	Prof. Dr. rer. nat. habil. Frank-Michael Schleif
Contact:	f.schleif@cs.bham.ac.uk
URL:	http://techfak.uni-bielefeld.de/∼ fschleif/mlr/mlr.html
ISSN:	1865-3960

1 Tenth Mittweida Workshop on Computational Intelligence

From 25 June to 27 June 2018 we had the pleasure to organize and attend the tenth Mittweida Workshop on Computational Intelligence (MiWoCi 2018). Multiple scientists from the University of Bielefeld, HTW Dresden, the University of Groningen (NL), the University of Birmingham (UK), the University of Applied Sciences Mittweida, the University of Applied Sciences Wuerzburg-Schweinfurt and the Porsche AG met in Mittweida, Germany, to continue the tradition of the Mittweida Workshops on Computational Intelligence - *MiWoCi'2018*.

The aim was to present their current research, discuss scientific questions, and exchange ideas. The seminar centered around topics in machine learning, signal processing and data analysis, covering fundamental theoretical aspects as well as recent applications, partially in the frame of innovative industrial cooperations. This volume contains a collection of abstracts which accompany some of the discussions and presented work of the MiWoCi Workshop.

Our particular thanks for a perfect local organization of the workshop go to Thomas Villmann as spiritus movens of the seminar and his PhD and Master students.

Mittweida, June, 2018
Frank-M. Schleif

[1]E-mail: frank-michael.schleif@fhws.de
[2]University of Appl. Sc. Wuerzburg-Schweinfurt, Wuerzburg, Germany

Learning Pharmacokinetic Models

Kerstin Bunte*

*University of Groningen, Groningen, NL

Abstract

To understand trends in individual responses to medication, one can take a purely data-driven machine learning approach, or alternatively apply pharmacokinetics combined with mixed-effects statistical modelling. To take advantage of the predictive power of machine learning and the explanatory power of pharmacokinetics, a latent variable mixture model for learning clusters of pharmacokinetic models is proposed and demonstrated on a clinical data set. The proposed strategy automatically constructs different population models that are not based on prior knowledge or experimental design, but result naturally as mixture component models of the global latent variable mixture model. The parameter of the underlying multi-compartment ordinary differential equation model are analyzed via identifiability analysis on the observable measurements, which reveals the model is structurally locally identifiable. Further approximation with a perturbation technique enables efficient training of the proposed probabilistic latent variable mixture clustering technique using Estimation Maximization.

Evaluation of Galaxy classification schemes with GMLVQ + feature learning for galaxy characterization: possible directions & open questions

Aleke Nolte*

*University of Groningen, Groningen, NL

Abstract

In Astronomy, automatic classification of galaxies is becoming increasingly important as astronomic surveys are generating more and more data. Yet, galaxy classification is not a well defined problem: Classification schemes are numerous and are commonly hand-designed, thereby possibly underling human cognitive biases. Building on work previously presented at ESANN, we investigate with the help of prototype-based methods how well a particular galaxy classification scheme is supported by the data. While our previous dataset contained only a handful of features, our current analysis is based on a variety of galaxy descriptors derived from photometric and spectroscopic observations. However, a potential problem of some of the galaxy descriptors is that they are based on historically grown model-assumptions which have been developed on the basis of bright and clearly-visible nearby galaxies, and may therefore not adequately describe fainter, or more distant galaxies. To explore possible alternative descriptors, we intend to also present ideas and open questions on feature learning for galaxy characterization, if time allows.

Learning Vector Quantization and its privacy

Johannes Brinkrolf

Bielefeld University, CITEC - Center of Excellence, Germany

Abstract

Digital information is collected daily in growing volumes. Mutual benefits drive the demand for the exchange and publication of data among parties. However, it is often unclear how to handle these data properly in the case that the data contains sensitive information. Differential privacy has become a powerful principle for privacy-preserving data analysis tasks in the last few years, since it entails a formal privacy guarantee for such settings. This is obtained by a separation of the utility of the database and the risk of an individual to lose his/her privacy.

In the workshop contribution we briefly review the problem of statistical disclosure control under differential privacy model and address the question how much the prototypes differ from those obtained from similar training sets. Furthermore, we present an approach for gaining a differential private LVQ model from learned models on different subsets via the sample and aggregate framework.

Towards Fair LVQ - Introducing Fairness Criteria into the GLVQ Cost Function

Astrid Bunge *[1], Carolin Hainke[1], Leon Sindelar[1], Matthias Vogelsang[1], Benjamin Paaßen[1], and Barbara Hammer[1]

[1]Machine Learning Group
Center of Excellence Cognitive Interaction Technology
Bielefeld University

Machine learning methods promise to speed up and ease human decision making in various fields, such as finance, jurisprudence and medicine. Yet, just like their human counter part, these decisions are prone to prejudice through biased data, resulting in possibly discriminatory decisions[2]. One approach to address such biases is to incorporate a formalized fairness criterion to the objective function of a machine learning algorithm. We have extended the error function of the generalized learning vector quantization classifier in [1] with the classic and normalized mean difference term referenced in [2]. The modification punishes any differential treatment between a protected group and its complement. By evaluating the effect of this fairness term on an artificial and real data set from the educational domain, we observed an increase in fairness under certain circumstances, while retaining most of the classification accuracy.

References

[1] Atsushi Sato and Keiji Yamada. "Generalized learning vector quantization". In: *NIPS'95 Proceedings of the 8th International Conference on Neural Information Processing Systems* (1995), pp. 423–429.

[2] I. Zliobaite. "Measuring discrimination in algorithmic decision making". In: *Data Mining and Knowledge Discovery* 31.4 (2017), pp. 1060–1098. DOI: https://doi.org/10.1007/s10618-017-0506-1.

*Corresponding author: abunge@techfak.uni-bielefeld.de

Multi-Label LVQ for Multi-Class Classification Learning

M. Kaden[1], A. Villmann[1,2] and , T. Villmann[1]

[1]Hochschule Mittweida, SICIM, CI-Group,
[2]Schulzentrum Döbeln-Mittweida

Classification learning usually deals with problems where data have to be assigned to one certain class. This scenario, however, frequently does not match with real world experiences, where objects/subjects may belong to several classes. For example, patients may suffer from multiple diseases or scientific articles share quite a few authors. For these examples, data sample might belong to more than one class (illness/author). In contrast, possibilistic classification deals with probabilistic class decisions and class belongings. Both approaches are known as multiple classification problems or multilabel classification [1].

In this contribution we address the multilabel classification problem for learning vector quantization models. Particularly, we discuss how to deal with multilabels for soft learning vector quantization introduced in [2]. Thereby, log-likelihood ratio or cross-entropy are possible loss functions [3]. Further, we discuss approaches which realize multilabel classification in generalized learning vector quantization (GLVQ,[4]). The first approach considers sets of best matching prototypes regarding the given multilabeled training sample whereas in a seconsd approach a cross entropy approach is used to model the possibility for more than one class assignments. The cross-entropy approach for GLVQ considers the classification of data samples as a probabilistic event keeping the idea of the classifier function [5].

For both algorithmic approaches, possibilistic as well as possibilistic strategies are discussed. Further, the problem of an adequate performance evaluation for those methods will be addressed during the talk.

References

[1] F. Herrera, F. Charte, A.J. Rivera, and M.J. del Jesus. *Multi-Label Classification – Problem Analysis, Metrics and Techniques*. Springer, 2016.

[2] S. Seo and K. Obermayer. Soft learning vector quantization. *Neural Computation*, 15:1589–1604, 2003.

[3] A. Villmann, M. Kaden, S. Saralajew, and T. Villmann. Probabilistic learning vector quantization with cross-entropy for probabilistic class assignments in classification learning. In L. Rutkowski, R. Scherer, M. Korytkowski, W. Pedrycz, R. Tadeusiewicz, and J.M. Zurada, editors, *Proceedings of the 17th International Conference on Artificial Intelligence and Soft Computing - ICAISC, Zakopane*, LNCS 10841, pages 736–749, Cham, 2018. Springer International Publishing, Switzerland.

[4] A. Sato and K. Yamada. Generalized learning vector quantization. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8. Proceedings of the 1995 Conference*, pages 423–9. MIT Press, Cambridge, MA, USA, 1996.

[5] A. Villmann, M. Kaden, S. Saralajew, W. Hermann, M. Biehl, and T. Villmann. Reliable patient classification in case of uncertain class labels using a cross-entropy approach. In M. Verleysen, editor, *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2018), Bruges (Belgium)*, pages 153–158, Louvain-La-Neuve, Belgium, 2018. i6doc.com.

Tree Edit Distance Learning with Median GLVQ and Symbol Embeddings

Benjamin Paaßen[1], Claudio Gallicchio[2], Alessio Micheli[2], and Barbara Hammer[1]

[1]Center of Excellence Cognitive Interaction Technology
Bielefeld University

[2]Department of Computer Science
University of Pisa

This contribution is based on the ICML 2018 Paper [6].

For vectorial data, metric learning has yielded tremendous improvements in classification accuracy and can be considered a standard method [1, 7]. Recent research has tried to translate these successes to structured data metrics, in particular edit distances [1, 5]. However, metric learning for edit distances is complicated my multiple challenges. First, efficint edit distance algorithms rely on metric conditions on the metric parameters [4], which are difficult to enforce during learning. Second, changing the metric parameters can also change the optimal edit scripts, making a direct optimization infeasible [5]. Finally, most indirect optimization approaches require frequent updates of all pairwise edit distances, making them slow for bigger data sets [2].

To address these challenges, we have developed a new metric learning approach for tree edit distance learning, which works as follows. First, we represent the data via few prototypes by means of median relational GLVQ [3]; second, we compute all cheapest edit scripts between data points and there closest correct and closest wrong prototypes using a novel forward-backward algorithm [4]; third, we optimize vectorial representations of the tree labels according to the GLVQ cost function. We iterate these three steps until convergence.

The use of a vectorial embedding ensures metric conditions, while the use of prototypes ensures that only a small number of backtraces needs to

be computed in each optimization step. This reflects in a favorable scaling behavior, making our new metric learning scheme applicable to data sets of thousands of trees and hundreds of thousands of nodes.

Our experimental results on one artificial and five real-world data sets show that our new metric learning scheme outperforms the state-of-the-art in tree edit distance metric learning and improves upon the standard tree edit distance in almost all cases.

References

[1] Aurélien Bellet, Amaury Habrard, and Marc Sebban. "A Survey on Metric Learning for Feature Vectors and Structured Data". In: *arXiv* abs/1306.6709 (2014). eprint: 1306.6709. URL: http://arxiv.org/abs/1306.6709.

[2] Aurélien Bellet, Amaury Habrard, and Marc Sebban. "Good edit similarity learning by loss minimization". In: *Machine Learning* 89.1 (Oct. 2012), pp. 5–35. DOI: 10.1007/s10994-012-5293-8.

[3] David Nebel et al. "Median variants of learning vector quantization for learning of dissimilarity data". In: *Neurocomputing* 169 (2015), pp. 295–305. DOI: 10.1016/j.neucom.2014.12.096.

[4] Benjamin Paaßen. "Revisiting the tree edit distance and its backtracing: A tutorial". In: *ArXiv e-prints* (2018). URL: https://arxiv.org/abs/1805.06869.

[5] Benjamin Paaßen, Bassam Mokbel, and Barbara Hammer. "Adaptive structure metrics for automated feedback provision in intelligent tutoring systems". In: *Neurocomputing* 192 (2016), pp. 3–13. DOI: 10.1016/j.neucom.2015.12.108.

[6] Benjamin Paaßen et al. "Tree Edit Distance Learning via Adaptive Symbol Embeddings". In: *Proceedings of the 35th International Conference on Machine Learning (ICML 2018)*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. accepted. Stockholm, 2018.

[7] Petra Schneider, Michael Biehl, and Barbara Hammer. "Adaptive Relevance Matrices in Learning Vector Quantization". In: *Neural Computation* 21.12 (2009), pp. 3532–3561. DOI: 10.1162/neco.2009.11-08-908.

Objective Feature Selection using GMLVQ with Directly Incorporated L_1-Regularization

Falko Lischke[1⋆], Thomas Neumann[1], Sven Hellbach[1], Thomas Villmann[2], and Hans-Joachim Böhme[1]

[1] University of Applied Sciences Dresden, Friedrich-List-Platz 1, 01069 Dresden, Germany {lischke,neumann,hellbach,boehme}@htw-dresden.de
[2] Saxony Institute for Computational Intelligence and Machine Learning, Univ. Applied Sciences Mittweida, 09648 Mittweida, Germany
thomas.villmann@hs-mittweida.de

Frequently, high-dimensional features are used to represent data to be classified. One such field of application with high-dimensional features is speech-based emotion recognition. The development of the feature sets used for these applications can be seen monitoring the respective changes in approaches presented for the Interspeech challenges since 2009. The 384 features used in the first Interspeech Emotion Challenge 2009 have become 6373 features since 2012 [8][9]. Linear SVM are still frequently used for classification, as in [10]. Instead of the predefined feature sets, more and more attempts are being made to have them learned automatically from artificial neural networks (MLP) [11]. In order to learn optimal features for MLP, it is assumed that the available data contains a large amount of variations provided by a huge database. In contrast, prototype-based methods frequently can work successfully with fewer data [1].

In our MiWoCI contribution we propose a new approach to learn interpretable classification models from such high-dimensional data representation. To this end, we extend a popular prototype-based classification algorithm, the matrix learning vector quantization (GMLVQ), to incorporate an enhanced feature selection objective via L_1-regularization [7]. In contrast to previous work, we propose a framework that directly optimizes this objective using the alternating direction method of multipliers (ADMM) and manifold optimization.

To incorporate the idea of feature selection into GMLVQ, we add a regularization term $R(\Omega) = \|\Omega\|_1$ to the GMLVQ optimization objective $E(\Omega, W, X)$

$$\underset{\Omega, W}{\text{minimize}}\ E(\Omega, W, X) + \xi R(\Omega)\ , \tag{1}$$

where $\xi > 0$ is a regularization parameter to control sparsity of Ω. We suggest for optimization the ADMM as a proximal algorithm to optimize Eq. 1 directly without approximation of the L_1-norm [3]. For the optimization using ADMM, we decouple the data and regularization term by incorporation of a second variable $\phi \in \mathbb{R}^{n \times n}$:

$$\begin{aligned} &\underset{\Omega, W, \phi}{\text{minimize}}\ E(\Omega, W) + \xi R(\phi) \\ &\text{subject to } \Omega = \phi,\ \|\Omega\|_2^2 = 1\ . \end{aligned} \tag{2}$$

⋆ This work was supported in part by SAB grant number 100231931.

Manifold optimization is used in the update steps due to constraints and simultaneous optimization of several variables. A detailed description of the approach can be found in [6].

We show that our method achieves state-of-the-art results on an artificial data set from Bojer et al. [2] and on the Berlin Database of Emotional speech [4] with eGeMAPS features [5] and show its abilities to select relevant dimensions from the features. In both experiments, GMLVQ with L_1-regularization based on our framework achieves similar accuracies as standard classifiers (Decision Tree and linear SVM). In addition, the accuracy could be increased by an additional regularization of the manually selected eGeMAPS features. This shows that an objective feature selection can result in higher accuracy than subjectively selected features. Our optimization framework offers an opportunity to select features from more extensive feature sets based on objective criteria.

References

1. Biehl, M., Hammer, B., Villmann, T.: Prototype-based models in machine learning. Wiley Interdisciplinary Reviews: Cognitive Science 7(2), 92–111 (2016)
2. Bojer, T., Hammer, B., Schunk, D., Von Toschanowitz, K.: Relevance determination in learning vector quantization. In: Proc. of ESANN (2001)
3. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning 3(1), 1–122 (2011)
4. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. In: Interspeech. vol. 5, pp. 1517–1520 (2005)
5. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE TAC 7(2), 190–202 (April 2016)
6. Lischke, F., Neumann, T., Hellbach, S., Villmann, T., Böhme, H.J.: Direct incorporation of l_1 -regularization into generalized matrix learning vector quantization. In: ICAISC. pp. 657–667. Springer (2018)
7. Schneider, P., Biehl, M., Hammer, B.: Adaptive relevance matrices in learning vector quantization. Neural Computation 21(12), 3532–3561 (2009)
8. Schuller, B., Steidl, S., Batliner, A.: The interspeech 2009 emotion challenge. In: 10th Annual Conference of the ISCA (2009)
9. Schuller, B., Steidl, S., Batliner, A., et al.: The interspeech 2017 computational paralinguistics challenge: Addressee, cold & snoring. In: ComParE, Interspeech 2017. pp. 3442–3446 (2017)
10. Sinith, M.S., Aswathi, E., Deepa, T.M., Shameema, C.P., Rajan, S.: Emotion recognition from audio signals using support vector machine. In: 2015 IEEE RAICS. pp. 139–144 (Dec 2015)
11. Wen, G., Li, H., Huang, J., Li, D., Xun, E.: Random deep belief networks for recognizing emotions from speech signals. Computational intelligence and neuroscience 2017 (2017)

Dropout in Learning Vector Quantization Networks for Regularized Learning and Classification Confidence Estimation

T. Villmann[1], J.R.D. John Ravichandran[1], S. Saralajew[2], and M. Biehl[3]

[1]University of Applied Sciences Mittweida [2]Dr. hc. F. Porsche AG Weissach [3]University Groningen

1 Introduction

Dropout during training of deep multilayer perceptron networks (deep MLP) is an appropriate method to prevent the network from overfitting [1]. Further, dropout during the working phase can be used to judge the confidence of the network regarding the output [2]. The output could be a regression value or a class label depending on the task. For single perceptrons with weights $\boldsymbol{\omega}_i \in \mathbb{R}^n$ and biases β_i the outputs for a given data vector $\mathbf{x} \in \mathbb{R}^n$ are calculated as $h_i(\mathbf{x}) = g_i(\langle \boldsymbol{\omega}_i, \mathbf{x} \rangle_E + \beta_i)$, where $\langle \boldsymbol{\omega}_i, \mathbf{x} \rangle_E$ is the Euclidean inner product. Dropout in antework of perceptrons then is realized by setting $\omega_{ij} = 0$ randomly with probability p_{drop}.

Learning vector quantization (LVQ) was not studied regarding dropout techniques so far. Yet, a respective investigation should be comparable to the approach in (multi-layer) perceptrons. For this purpose, we consider in this contribution the matrix learning LVQ variant (GMLVQ,[3]) and relate this model to a multilayer network structure comparable to MLP. We denote the resulting model as LVQ-multylayer-network LVQ-MLN.

2 The LVQ-MLN model

2.1 Model Description

Standard GMLVQ uses the distance measure $\tilde{\delta}_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{w}_k) = (\mathbf{\Omega}(\mathbf{x} - \mathbf{w}_k))^2$ for similarity between data $\mathbf{x}$ and prototypes $\mathbf{w}_k$ where $\mathbf{\Omega} \in \mathbb{R}^{m \times n}$ is a projection matrix. An alternative for $\tilde{\delta}_{\mathbf{\Omega}}$ is the measure

$$d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{w}_k) = (\mathbf{\Omega}\mathbf{x} - \mathbf{w}_k)^2 \tag{1}$$

where

$$\mathbf{\Omega}\mathbf{x} = (\langle \boldsymbol{\omega}_1, \mathbf{x} \rangle_E, \ldots, \langle \boldsymbol{\omega}_m, \mathbf{x} \rangle_E) \tag{2}$$

and $\boldsymbol{\omega}_j$ are the row vectors of $\mathbf{\Omega}$. Now, the prototypes live in the projection space $\mathbb{R}^m$. The GMLVQ network realizes the class assignment

$$c(\mathbf{x}) = c\left(\mathbf{w}_{s(\mathbf{x})}\right)$$

for a data sample $\mathbf{x}$ by means of a winner-take-all competition (WTAC)

$$s(\mathbf{x}) = \operatorname{argmin}_k \left(d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{w}_k)\right) \tag{3}$$

where $c(\mathbf{w}_k)$ is the class label of prototype $\mathbf{w}_k$.

Now we consider a GMLVQ network as a multilayer network containing two hidden layers $\mathbf{h}^{\mathrm{I}}$ and $\mathbf{h}^{\mathrm{II}}$ as suggested in [4], see Fig. 1.

The nodes h_i^{I} of the first hidden layer $\mathbf{h}^{\mathrm{I}} \in \mathbb{R}^{n_p}$ are perceptron units according to

$$h_i^{\mathrm{I}}(\mathbf{x}) = g_i^{\mathrm{I}}\left(\langle \boldsymbol{\omega}_i, \mathbf{x} \rangle_E + \beta_i^{\mathrm{I}}\right) \tag{4}$$

with activation functions g_i^{I}, perceptron weight vectors $\boldsymbol{\omega}_i \in \mathbb{R}^n$ and biases $\beta_i^{\mathrm{I}} \in \mathbb{R}^n$. Thus, the first layer performs a maybe nonlinear projection

$$\mathbf{h}^{\mathrm{I}}(\mathbf{x}) = \mathbf{g}^{\mathrm{I}}_{\mathbf{\Omega},\boldsymbol{\beta}}(\mathbf{x}) \tag{5}$$

of the data depending on the activation functions $\mathbf{g}^{\mathrm{I}}_{\mathbf{\Omega},\boldsymbol{\beta}}$ with $\mathbf{\Omega} = \left(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_{n_p}\right)$ and the bias vector $\boldsymbol{\beta} \in \mathbb{R}^{n_p}$. Therefore, this layer is denoted as projection layer in this context. If $\beta_i^{\mathrm{I}} = 0$ and $g_i^{\mathrm{I}}(z) = \mathrm{id}(z)$ is the identity for all $i = 1 \ldots n_p$ the projection simply becomes $\mathbf{h}^{\mathrm{I}}(\mathbf{x}) = \mathbf{\Omega}\mathbf{x}$ as used in (1)

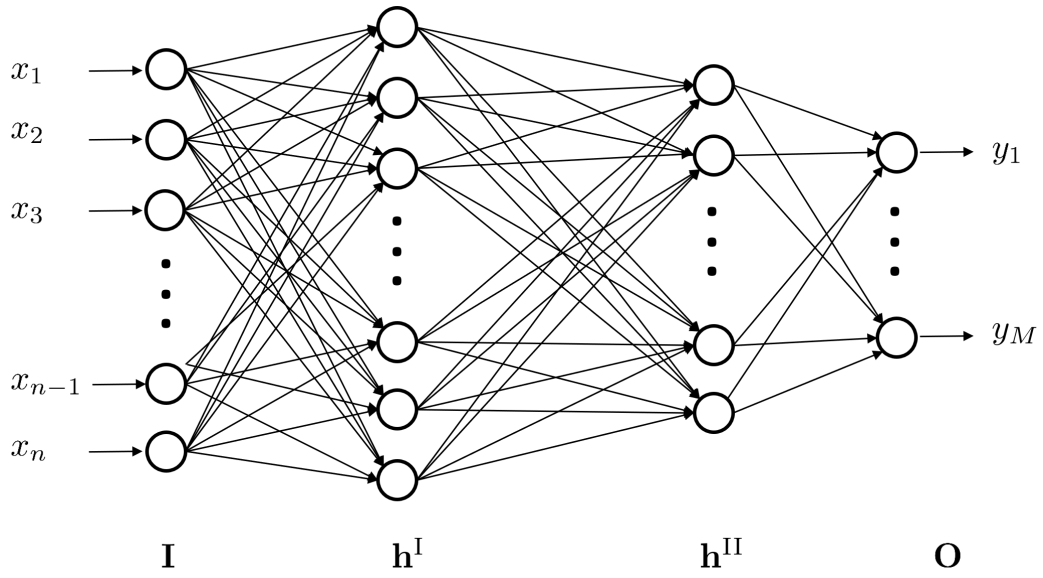


Figure 1: Illustration of an LVQ-MLP network with two hidden layers.

The second layer $\mathbf{h}^{\mathrm{II}}$ is fully connected to the previous layer $\mathbf{h}^{\mathrm{I}}$ via

$$h_j^{\mathrm{II}}(\mathbf{x}) = g^{\mathrm{II}}\left(d\left(\mathbf{h}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_j\right)\right) \tag{6}$$

realizing the prototype response. Here, d is an arbitrary (differentiable) dissimilarity measure and g^{II} is the activation function for the second layer usually chosen as the identity function $\mathrm{id}(z) = z$. If d is the squared Euclidean metric, $d\left(\mathbf{h}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_j\right) = d_{\mathbf{\Omega}}(\mathbf{x}, \mathbf{w}_k)$ is valid.

For a crisp classifier network, the output layer $\mathbf{O} \in \mathbb{R}^M$ is calculated as

$$O_l = \sum_{k=1}^{M} H\left(h_l^{\mathrm{II}}(\mathbf{x}) - h_k^{\mathrm{II}}(\mathbf{x})\right) \tag{7}$$

where

$$H(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ & is \\ 0 & \text{else} \end{cases}$$

the Heaviside function. Thus, O_l returns the winning rank of the prototype $\mathbf{w}_l$. Hence, $O_l = 1$ is valid iff $l = s(\mathbf{x})$ with

$$s(\mathbf{x}) = \mathrm{argmin}_k\left(h_k^{\mathrm{II}}(\mathbf{x})\right) \tag{8}$$

realizing the WTAC (3). Therefore, we denote the output layer also as competition layer. Finally, the data point $\mathbf{x}$ is assigned to the class of the corresponding winning output unit $c\left(\mathbf{w}_{s(\mathbf{x})}\right)$. Thereby, the formula (7) for the determination of the winning rank is equivalent to the winning rank determination known from the neural gas network [5].

2.2 The Loss Function of LVQ-MLN

The loss function for the LVQ-MLN is based on the output calculation according to (7). Let $h_+^{\mathrm{II}}(\mathbf{x})$ be defined as $h_+^{\mathrm{II}}(\mathbf{x}) = h_{s_+}^{\mathrm{II}}(\mathbf{x})$ with

$$s_+ = \arg\min_k \left\{O_k | c(\mathbf{w}_k) = c(\mathbf{x})\right\}$$

and $h_-^{\mathrm{II}}(\mathbf{x}) = h_{s_-}^{\mathrm{II}}(\mathbf{x})$ with

$$s_- = \arg\min_k \left\{O_k | c(\mathbf{w}_k) \neq c(\mathbf{x})\right\}$$

indicating the best correct and best incorrect classifying prototypes according to the output layer. Then the local loss is given as

$$L(\mathbf{x}, W, \boldsymbol{\Omega}, \boldsymbol{\beta}) = \phi_{\theta,\vartheta}\left(\mu\left(\mathbf{x}_k, \mathbf{h}^{\mathrm{II}}\right)\right) \tag{9}$$

with

$$\mu\left(\mathbf{x}_k, \mathbf{h}^{\mathrm{II}}\right) = \frac{h_+^{\mathrm{II}}(\mathbf{x}) - h_-^{\mathrm{II}}(\mathbf{x})}{h_+^{\mathrm{II}}(\mathbf{x}) + h_-^{\mathrm{II}}(\mathbf{x})}$$

is the equivalent to the classifier function of GMLVQ and

$$\phi_{\theta,\vartheta}(z) = \frac{1}{1 + \exp\left(\frac{z}{\theta} - \vartheta\right)}\,. \tag{10}$$

is the sigmoid function known from GMLVQ. Remember, the layer $\mathbf{h}^{\mathrm{II}}(\mathbf{x})$ is connected to layer $\mathbf{h}^{\mathrm{I}}(\mathbf{x})$ via (6) and $\mathbf{h}^{\mathrm{I}}(\mathbf{x})$ depends on the matrix $\boldsymbol{\Omega}$ by the projection layer (5) and the perceptron layer (4). Applying these replacements we obtain

$$L(\mathbf{x}, W, \boldsymbol{\Omega}, \boldsymbol{\beta}) = \phi_{\theta,\vartheta}\left(\frac{d\left(\mathbf{g}_{\boldsymbol{\Omega},\boldsymbol{\beta}}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_{s_+}\right) - d\left(\mathbf{g}_{\boldsymbol{\Omega},\boldsymbol{\beta}}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_{s_-}\right)}{d\left(\mathbf{g}_{\boldsymbol{\Omega},\boldsymbol{\beta}}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_{s_+}\right) + d\left(\mathbf{g}_{\boldsymbol{\Omega},\boldsymbol{\beta}}^{\mathrm{I}}(\mathbf{x}), \mathbf{w}_{s_-}\right)}\right) \tag{11}$$

where we have taken $g^{\mathrm{II}}(z) = \mathrm{id}(z)$.

Learning in LVQ-MLN can be performed by stochastic gradient descent learning (SGDL,[6, 7]) as usual in multilayer network learning [8].

2.3 Dropout in LVQ-MLN network

The dropout strategy in the LVQ-MLN can easily be realized applying them to the matrix $\mathbf{\Omega}$ of the projection layer $\mathbf{g}^{\mathrm{I}}_{\mathbf{\Omega},\boldsymbol{\beta}}(\mathbf{x})$ from (5). This could be done during training preventing overfitting or for confidence estimation when applied during the test phase. The training dropout should be compared with other regularization techniques as known for GMLVQ [9] whereas confidence considerations should be compared with reject option methods [10] or conformal prediction analysis for LVQ [11, 12]. Obviously, LVQ-MLN offers great similarity to MLP networks, and therefore, a comparison with MLP-classifiers is mandatory.

References

[1] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[2] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In M.F. Balcan and K.Q. Weinberger, editors, *Proceedings of the International Conference on Machine Learning, New York, New York, USA*, volume 48, pages 1050–1059, 2016.

[3] P. Schneider, B. Hammer, and M. Biehl. Adaptive relevance matrices in learning vector quantization. *Neural Computation*, 21:3532–3561, 2009.

[4] T. Villmann, M. Biehl, A. Villmann, and S. Saralajew. Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning. In *Proceedings of the 12th Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM2017+)*, pages 248–255. IEEE Press, 2017.

[5] Thomas M. Martinetz, Stanislav G. Berkovich, and Klaus J. Schulten. 'Neural-gas' network for vector quantization and its application to time-series prediction. *IEEE Trans. on Neural Networks*, 4(4):558–569, 1993.

[6] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22:400–407, 1951.

[7] S. Graf and H. Lushgy. *Foundations of Quantization for Probability Distributions*, volume 1730 of *Lect. Notes in Mathematics*. Springer, Berlin, 2000.

[8] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[9] P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and Michael Biehl. Regularization in matrix relevance learning. *IEEE Transactions on Neural Networks*, 21(5):831–840, 2010.

[10] L. Fischer, D. Nebel, T. Villmann, B. Hammer, and H. Wersing. Rejection strategies for learning vector quantization a comparison of probabilistic and deterministic approaches. In T. Villmann, F.-M. Schleif, M. Kaden, and M. Lange, editors, *Advances in Self-Organizing Maps and Learning Vector Quantization: Proceedings of 10th International Workshop WSOM 2014, Mittweida*, volume 295 of *Advances in Intelligent Systems and Computing*, pages 109–118, Berlin, 2014. Springer.

[11] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[12] F.-M. Schleif, X. Zhu, and B. Hammer. A conformal classifier for dissimilarity data. In L. Iliadisand I. Maglogiannis, H. Papadopoulos, K. Karatzas, and S. Siouta, editors, *Proceedings of AIAI 2012, Halkidiki, Greece*, volume 382 of *IFIP Advances in Information and Communication Technology*, pages 234–243, Berlin, 2012. Springer.

A comparison of classifier learning strategies and their counterparts in adaptive filter theory

Daniel Staps†, Alexander Lampe† and Julia Schulte‡
† Hochschule Mittweida, Germany, {dstaps, lampe}@hsmw.de
‡ CI Tech Sensors AG, Burgdorf, Switzerland, julia.schulte@citechsensors.com

The requirements on the reliability and quality of security paper processing have been rising continuously in recent years. Focusing on the field of banknote processing in banknote readers, this results in a steady improvement of existing and the development of new image processing algorithms. In the majority of cases, the target of those algorithms is to distinguish a limited number of banknote classes ranging from 2 to about 100 and the classification rules are based on reference samples for each class. The number of reference samples for each class varies between a few, e.g., 10, samples, up to a huge number, e.g., 10000, samples. In view of the last figure, the relatively small number of banknote classes to be discriminated and the capabilities of machine learning algorithms, the question arose whether the latter can be trained such that they achieve a classification performance comparable to classical banknote processing algorithms, i.e., fulfilling existing industry standards.

As to provide a first answer, a tensorflow framework based on a convolutional neural network (CNN) inception v3 model [1, 2] was chosen and its last layers, the classification part, were trained to detect the different denominations of EUR and USD banknotes. This training was done with different learning strategies, with varying parameters, e.g. learning rate and training batch size, and with different banknote image formats being fed into the network. Measuring the resulting classification accuracies, the investigation shows that with the chosen network a performance as good as that of current signal processing algorithms can be achieved in principle for valid banknote records. However, the resulting performance heavily relies on the learning strategy chosen and the selected parameters. The detailed results will be presented and discussed at the conference.

In addition, when trying to understand the details and differences of the learning strategies applied for optimization of the classifier's weights, it turns out that the underlying algorithms are quite similar to those used for optimization of adaptive filters' coefficients, e.g., the ADAM algorithm [3] resembles the NLMS algorithm [4]. Consequently, the characteristics of different optimization strategies used in machine learning can in part be deduced from those of their adaptive filter counterparts.

References

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Jun. 2014, pp. 1701–1708.
[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions."
[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization."
[4] F. Yassa, "Optimality in the choice of the convergence factor for gradient-based adaptive algorithms," *and Signal Processing IEEE Transactions on Acoustics, Speech*, vol. 35, no. 1, pp. 48–59, Jan. 1987.

The statistical physics of learning in a nutshell - news from the stoneage of machine learning

Michael Biehl*

*University of Groningen, Groningen, NL

Abstract

We revisit the successful statistical physics of learning which has contributed significantly to the theory of neural networks and machine learning. In this framework, large systems with many adaptive parameters are considered which are optimized by means of stochastic training processes. The formal treatment of learning systems in thermal equilibrium situtations facilitates the application of methods borrowed from statistical physics and provides mathematically exact descriptions of, for instance, typical learning curves in model situations. We review the basic concepts in terms of the perceptron classifier and layered neural networks for regression tasks. Finally, we discuss potential applications of the framework in view of the recently regained popularity of neural networks.

Processing Gene Expression Data for Detection of mRNA Degradation Patterns by Cluster Analysis Using a Bio-specific Similarity Measure

Katrin Bohnsack[1], Röbbe Wünschiers[1], and Thomas Villmann[2]

[1]University of Applied Sciences Mittweida, Research Group Biotechnology/Chemistry,
[2]University of Applied Sciences Mittweida, Saxony Institute for Computational Intelligence and Machine Learning

In this work, the task is to cluster microarray gene expression data of the cyanobacterium *Nostoc* PCC 7120 for detection of messenger RNA (mRNA) degradation patterns. We search for characteristic patterns of degradation which are caused by specific enzymes (ribonucleases) allowing a further biological investigation regarding biochemical mechanisms behind.

The mRNA degradation is part of the regulation of gene expression because it regulates the amount and the longlivity of mRNA, which is available for translation into proteins. A particular class of RNA degrading enzymes are exoribonucleases which degrade the molecule from its ends, whereby a degradation from the 5' end, the 3' end or from both ends is theoretically possible [1, 2].

In this investigation, the information about exoribonucleolytic degradation is given in a microarray data set containing gene expression values of 1251 genes. The data set provides gene expression vectors containing the expression values of up to 10 short distinct sections of a gene ordered from the genes 5' end to its 3' end [3]. For each gene, two expression vectors are available for both nitrogen fixing and non-nitrogen fixing conditions, which have to be considered separately due to biological reasons. Accordingly, after filtering and preprocessing, two datasets for clustering are obtained consisting of 133 10-dimensional expression vectors. The preprocessing of data is described in detail in [4].

The similarity of the expression vectors is frequently judged by the Euclidean distance d_E or by the Spearman rank correlation ρ_S. Unfortunately, the rank correlation is not a similarity measure. Yet, the shifted value $\rho_S + 1$ would deliver a similarity [5]. However, due to the usually noisy expression values, small positive correlations might contain little to no corellation information. Thus, we recommend to apply a non-linear transormation of ρ_S according to

$$s_S(\rho_S, \beta, \theta) = \frac{1}{1 + \exp\left(-\frac{\rho_S - \beta}{\theta}\right)} \tag{1}$$

to obtain a dissimilarity value $d_S = 1 - s_S$. Thereby, the values for β and θ are in relation according to

$$\beta = \theta \cdot \ln\left(\frac{1-y}{y}\right) + x \tag{2}$$

whereby a fixed $x \in (-1, 1)$ is required to be mapped onto a given $y = s_s(\rho_S, \beta, \theta) \in (0, 1)$. Thus, a user specific differentiation between negative and positive correlated gene expression vectors is possible. Further, the choice of the values x and y allows an adequate adjustment regarding the noise level of gene expression values. After systematic evaluation, we have chosen $x = y = 0.5$ and $\theta = 0.05$ for our clustering experiments.

Clustering was performed using affinity propagation (AP,[6]). The number of clusters obtained by AP depends on the so-called self-similarity for the data vectors. This dependence was used to identify stable cluster soltuions by self-similarity control.

To evaluate the clustering results, several cluster validity measures are applied. Further, visual data inspections by t-SNE [7] as well as respective cluster visualizations are provided for interpretation analysis of clusters.

References

[1] J. Houseley and D. Tollervey. The many pathways of RNA degradation. *Cell*, 136:763–776, 2009.

[2] V.R. Kaberdin, D. Singh, and S. Lin-Chao. Composition and conservation of the mRNA-degrading machinery in bacteria. *Journal of Biomedical Sciences*, 18(23):1–12, 2011.

[3] S. Motameny and R. Wünschiers. Clustering approach to detect mRNA - degradation patterns from DNA-microarray gene-expression data. *Biosystems and Information Technology*, 1(1):6–13, 2012.

[4] K. Bohnsack. Clustering of gene expression data to detect mRNA degradation patterns. Technical Report in prep., University of Applied Sciences Mittweida, 2018.

[5] D. Nebel, M. Kaden, A. Bohnsack, and T. Villmann. Types of (dis−)similarities and adaptive mixtures thereof for improved classification learning. *Neurocomputing*, 268:42–54, 2017.

[6] B.J. Frey and D. Dueck. Clustering by message passing between data points. *Science*, 315:972–976, 2007.

[7] L. v.d.Maaten and G. Hinten. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Machine learning in biomedical datasets

Sreejita Ghosh

To work with Biomedical data one has to deal with issues associated with it, namely, heterogeneous measurements, imbalanced classes, and missing data. Our initial experiments with a variant of learning vector quantization (LVQ) that is capable of dealing with missing values (angle LVQ) gave us promising results. So we wanted to compare the performance of angleLVQ with a strategy one can follow when confronted with missing values. Generative modelling is one such strategy. Furthermore, LDA is very close to LVQ. Therefore we applied B.Marlin's generative linear discriminant analysis (LDA) [1] on the same datasets to compare its performance with that of angleLVQ. Additionally in our recent experiments we tried to compare 1) the effects of Euclidean and cosine distances, and 2) the effect of a hyper-parameter of angleLVQ, in classifying data (both synthetic and real) with missing values.

References

[1] Benjamin Marlin. *Missing data problems in machine learning.* PhD thesis, 2008.

Using GANs for dense three dimensional reconstruction of neuronal tissue from electron microscopy stacks

T. Bullmann, S. Oba, and S. Ishii

Department of Systems Science, Graduate School of Informatics, Kyoto University, Japan

June 25, 2018

Focused ion beam milling, combined with scanning electron microscopy (FIB-SEM), can be used to generate serial images through substantial volumes of neuropile, making it possible to capture subtle changes in spine structure as well as quantifying the local connectivity of several dendrites with passing axons. From detailed reconstruction of the neuropile, it is possible to recover many aspects of synaptic calcium signaling, which is involved in plasticity of synapses, thus making predictions for future in vivo imaging experiments. However, conventional reconstruction methods require skillful and time-consuming manual annotation.

To this end, various machine learning techniques have been used to achieve automatic or semi-automatic segmentation with minimal human annotation. Recently, deep convolutional networks have shown superhuman performance in automatic membrane segmentation, but they require huge amounts of labelled data and it is difficult to used trained classifiers on images obtained at different imaging conditions/species. We will present preliminary results of our attempt to use generative adversarial networks (GANs) for data augmentation of limited amount of labelled data for segmentation and reconstruction of the Drosophila neuropile.

Transfer Learning for Robust Control of Bionic Prostheses

Alexander Schulz Benjamin Paaßen Barbara Hammer

The aim of transfer learning [2] is to make use of knowledge from existing models in new domains and thereby avoid to train an entirely new model. This methodology is particularly promising if the trained model is complex but the relationship between the old and the new domain is simple, for example an approximately linear function. Recently, the framework of linear supervised transfer learning has been proposed which learns a mapping from the target to the source domain with the goal that the original model becomes applicable in the target domain [1, 3].

Here, we present a more efficient variant of linear supervised transfer learning for correcting electrode shifts in upper limb prosthesis control. For this purpose, we introduce a bias/restriction on the transfer mapping which reduces the number of parameters that need to be estimated to one. We evaluate our approach in a virtual grasping environment with a group of transradial amputees and a group of able-bodied subjects.

References

[1] B. Paaßen, A. Schulz, J. Hahne, and B. Hammer. *Neurocomputing*, 298:122 – 133, 2018.

[2] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, Oct 2010.

[3] S. Saralajew and T. Villmann. Transfer learning in classification based on manifold models and its relation to tangent metric learning. In C. J. Yoonsuck Choe, editor, *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN 2017)*, Anchorage, AK, USA, 2017. in press.

Handling Concept Drift and Domain Differences in an Online Learning Environment

Christoph Raab

June 21, 2018

Abstract

Supervised classification has a broad range of applications in different domains of interest. Typical, classification algorithms are batch based and processing data multiple times, building some model to represent class distributions and being able to predict labels of unseen data.

However, streaming applications producing a vast amount of data resulting in datasets, which are inefficient to learn by above approaches. These streams are either too large to fit in memory or processing in batch mode is not a constructive strategy due to constant arriving of new samples.[4]

Algorithms based on online learning are able to learn and predict data on the fly, processing data only once and, therefore, are able to operate on data streams. But, changing the learning technique is not sufficient to classify streams with high accuracy. A reason is the event of concept drift, which is a not negligible change in class distributions between two points in time.[5]

Ensemble approaches trying to tackle these issues and are build on top of an online classifier, but expanding the complexity of these algorithms further [5].

This results in less interpretable models because of nested bagging and restricts the underlying classification technique to the ensemble setting. This makes ensemble algorithms less interchangeable and, therefore, are hard to apply to online classifiers, which are not suitable to certain restrictions caused by ensemble techniques.

In this talk we, will discuss the above problems of concept drift handling, memory management, and interpretability by introducing current solutions [1][2][3], giving insights in current research potential.

References

[1] Vahida Attar, Pradeep Sinha, and Kapil Wankhade. A fast and light classifier for data streams. *Evolving Systems*, 1(3):199–207, 2010.

[2] Albert Bifet and Ricard Gavaldà. Learning from Time-Changing Data with Adaptive Windowing. *Proceedings of the 2007 SIAM International Conference on Data Mining*, pages 443–448, 2007.

[3] Pedro M Domingos and Geoff Hulten. Mining high-speed data streams. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, August 20-23, 2000*, pages 71–80, 2000.

[4] Sudipto Guha, Adam Meyerson, Nina Mishra, Rajeev Motwani, and Liadan O'Callaghan. Clustering Data Streams: Theory and Practice. *IEEE Trans. Knowl. Data Eng.*, 15(3):515–528, 2003.

[5] Vincent Lemaire, Christophe Salperwyck, and Alexis Bondu. *A Survey on Supervised Classification on Data Streams*, pages 88–125. Springer International Publishing, Cham, 2015.

Top Tier Conferences - What Makes the Difference Between Accept and Reject - Some Reviewer Insights

Frank-Michael Schleif*

*University of Applied Sciences Wuerzburg-Schweinfurt, Department of Computer Science, Wuerzburg, DE

Abstract

Conferences like NIPS , AISTAT, IJCAI, ICML, ECML, AAAI, COLT ... are very attractive for many researchers and to get a paper accepted can give a career a big push. The talk will give some insights, statistical information and fun facts on common patterns of good and accepted papers at the respective conferences as noticed by the presenter in the last years.

Visualizing classifiers of proximity data

Sascha Schleef[1,*], Alexander Schulz[1], Barbara Hammer[1]

[1]CITEC, Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany

[*]Corresponding author: sschleef@techfak.uni-bielefeld.de

Proximity data can be classified by different classifiers, such as LVQ or SVM based approaches, taking the kernel directly as input. But visualizing these classifiers is not always straight forward and the approach proposed in [2] is not applicable because the original data has no vector representation which can be interpolated.

For proximity data we can apply distance based methods for dimensionality reduction such as kernel t-SNE or kernel Fisher-t-SNE [1]. To visualize an applied classifier, like a kernel-SVM on such data, the approach is to calculate an implicit back projection into a hypothetical high dimensional space by also minimizing the Fisher-distance such that the proximity data can be interpreted as a scalar product in this space. This opens the way to calculate a similar back projection, like for vector data, in kernel space.

This approach is evaluated for visualizing kernel SVMs for a Fisher distance based t-SNE dimension reduction method [1] by comparing the original classifier with the visualized classifier.

References

[1] Alexander Schulz, Johannes Brinkrolf, and Barbara Hammer. Efficient Kernelization of Discriminative Dimensionality Reduction. *Neurocomputing*, 268(SI):34–41, 2017.

[2] Alexander Schulz, Andrej Gisbrecht, and Barbara Hammer. Using Discriminative Dimensionality Reduction to Visualize Classifiers. *Neural Processing Letters*, 42(1):27–54, 2015.

Entropy based evaluation measures for clustering and classification

Tina Geweniger and Thomas Villmann

The comparison of cluster or classification models with ground truth data or other models is usually done by the statistical evaluation of respective confusion matrices [1]. But sometimes such classical evaluation measures like accuracy and Kappa value are misleading and not conveying the full informational content. An example thereof is given in [3]. There, an information theoretic approach resulting in two scores based on the Shannon entropy and mutual information or conditional Shannon entropy is proposed. Thereby the (normalized) assignments contained in the confusion matrix are assumed to be joint probabilities allowing to calculate the scores by means of probability and conditional probability functions. For the special case of Shannon entropy the two score are identical.

Yet it is known that the numerical computation of measures based on the Shannon entropy is unstable for very small probabilities due to the logarithmic function inside the sum [2]. We proposed more robust alternative measures based on either Renyi or Tsallis entropy in [4]. Unfortunately, an essential property regarding the mutual information is not valid for these entropies. Therefore the two scores are no longer identical and different definitions for the conditional entropies have to be taken into account.

In our current contribution we will show the difficulties dealing with nonadditive entropies like the Tsallis entropy. Care has to be taken to assure symmetry in general and validity in case of dependend variables. Different scenarios will be considered to derive the evaluation scores and to point out the challenges.

References

[1] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[2] O. Onicescu. Theorie de l'information energie informationelle. In Tome, editor, *C. R. Acad. Sci.*, volume 263 of *A-B*, pages 841–842, 1966.

[3] E.K. Kao R.S. Holt, P.A. Mastromarino and M.B. Hurley. Information theoretic approach for performance evaluation of multi-class assignment systems. In *SPIE Defense, Security, and Sensing (Orlando)*, volume SPIE 7697, pages 1–12. SPIE The International Society for Optical Engineering, MIT Press, 2010.

[4] Thomas Villmann and Tina Geweniger. Multi-class and cluster evaluation measures based on renyi and tsallis entropies and mutual information. In Leszek Rutkowski et. al, editor, *Artificial Intelligence and Soft Computing*, LNAI 10841, page 736 ff., 2018.

Detection of noisy multi-manifold

Mohammad Mohammadi*

*University of Groningen, Groningen, NL

Abstract

A common assumption in ML is:

- Even for high dimensional data, they lie on a low dimensional manifold.

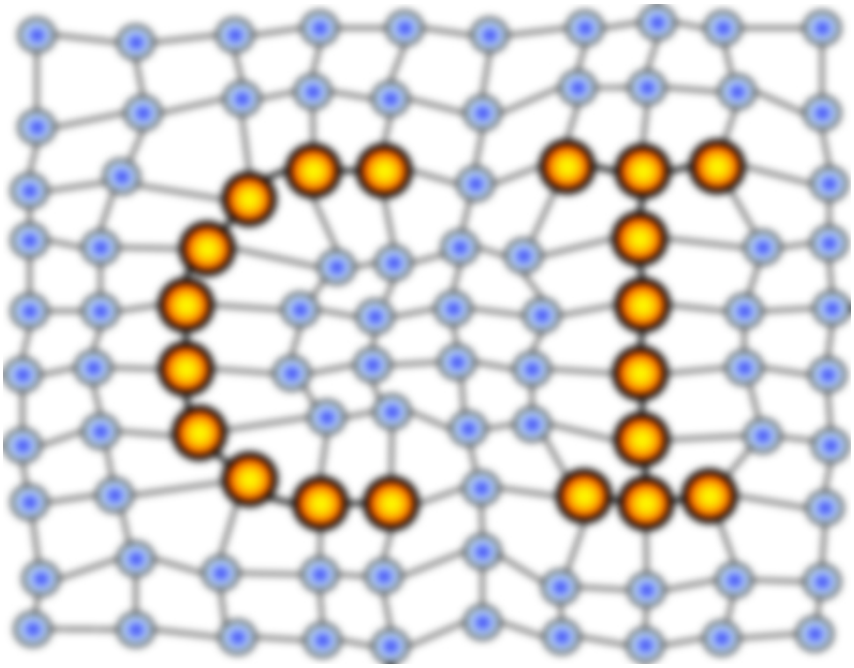
However, the noise can increase the dimensionality of the manifold. So the question arises:

- How to remove the points which dont belong to manifolds?

If we could find the answer, we could reconstruct the manifolds which is so helpful for the next actions. Here, we use a nature inspired approach (Ant colony) to recover the manifold.

MACHINE LEARNING REPORTS

Report 01/2018



Impressum

Machine Learning Reports ISSN: 1865-3960

▽ Publisher/Editors

Prof. Dr. rer. nat. Thomas Villmann
University of Applied Sciences Mittweida
Technikumplatz 17, 09648 Mittweida, Germany
• http://www.mni.hs-mittweida.de/

Prof. Dr. rer. nat. Frank-Michael Schleif
University of Birmingham
Edgbaston, B15 2TT Birmingham, UK,
• www.cs.bham.ac.uk/∼schleify/

▽ Copyright & Licence

Copyright of the articles remains to the authors.

▽ Acknowledgments

We would like to thank the reviewers for their time and patience.